University of Glasgow at WebCLEF 2005: Experiments in Per-Field Normalisation and Language Specific Stemming

Craig Macdonald, Vassilis Plachouras, Ben He,
Christina Lioma, and Iadh Ounis

University of Glasgow, G12 8QQ, UK
{craigm, vassilis, ben, xristina, ounis}@dcs.gla.ac.uk

Abstract. We participated in the WebCLEF 2005 monolingual task. In this task, a search system aims to retrieve relevant documents from a multilingual corpus of Web documents from Web sites of European governments. Both the documents and the queries are written in a wide range of European languages. A challenge in this setting is to detect the language of documents and topics, and to process them appropriately. We develop a language specific technique for applying the correct stemming approach, as well as for removing the correct stopwords from the queries. We represent documents using three fields, namely content, title, and anchor text of incoming hyperlinks. We use a technique called per-field normalisation, which extends the Divergence From Randomness (DFR) framework, to normalise the term frequencies, and to combine them across the three fields. We also employ the length of the URL path of Web documents. The ranking is based on combinations of both the language specific stemming, if applied, and the per-field normalisation. We use our Terrier platform for all our experiments. The overall performance of our techniques is outstanding, achieving the overall top four performing runs, as well as the top performing run without metadata in the monolingual task. The best run only uses per-field normalisation, without applying stemming.

1 Introduction

One of the main problems when applying language dependent retrieval techniques to multilingual collections is to identify the language in which the documents are written, in order to apply the appropriate linguistic processing techniques, such as stemming, or the removal of the appropriate stopwords. The EuroGOV collection used in WebCLEF 2005 is a crawl of European government Web sites, and includes documents written in a broad range of European languages [13]. Similarly, the used topics in WebCLEF 2005 are provided in any one of 11 different languages, as well as translated in English. We investigate a language specific stemming approach to deal with the multilingual setting. Our approach is based on identifying the language of the documents and the queries, and applying the appropriate stemmer. Our main hypothesis is that applying the

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 898–907, 2006.

appropriate stemmer for the language of each document and topic will increase the retrieval effectiveness of the search engine.

The WebCLEF 2005 monolingual task is a typical Web known-item finding search task, consisting of home page and named page finding queries, where there is only one relevant document, and the query corresponds to the name of this document. The evaluation results for similar tasks in the context of the Web track of TREC 2003 [3] and TREC 2004 [4] have shown that the anchor text and title of Web documents are effective sources of evidence for performing retrieval. The URL of Web documents is also very effective for identifying the home pages of Web sites. Therefore, we use the same Web Information Retrieval (IR) techniques in our WebCLEF 2005 participation.

We represent a Web document with three different fields, namely its content, title and the anchor text of its incoming hyperlinks. To apply an appropriate term frequency normalisation, and to combine the information from the different fields, we employ a technique called *per-field normalisation*, which extends *Normalisation 2*, a term frequency normalisation technique from the Divergence From Randomness (DFR) framework [1]. We also use evidence from the length of the document URL path in order to identify relevant home pages.

In all our submitted runs we use per-field normalisation and evidence from the document URL path. Depending on the method used to identify the language of documents and queries, we test various approaches for performing stemming in a robust and appropriate way on the tested multilingual setting. We use our Terrier IR platform [10] to conduct all the experiments.

Our results show that all our submitted runs to WebCLEF 2005 achieved outstanding retrieval performance. In particular, we achieved the overall top four performing runs, as well as the top performing run without the use of metadata in the monolingual task. The results suggest that the per-field normalisation seems to be effective in enhancing the retrieval performance, while there is less benefit from using the URLs of Web documents. In addition, our results suggest that the application of language specific stemming achieves good performance when the language of documents is identified correctly. However, the highest retrieval performance is achieved when no stemming is applied.

The remainder of this paper is organised as follows: Section 2 provides details of our methodology for language specific stemming, per-field normalisation, and evidence from the document URL path. Section 3 describes the experimental setting and our submitted runs to the monolingual task of WebCLEF 2005. We present and discuss the obtained results in Section 4. This paper closes with some concluding remarks in Section 5.

2 Searching a Multilingual Web Test Collection

We describe our proposed techniques for effectively searching a multilingual Web collection for known-item finding queries. As mentioned in the introduction, one major problem with multilingual test collections is how to apply the appropriate linguistic processing techniques, such as stemming, or the removal of the

appropriate stopwords. Section 2.1 presents our proposed language specific stemming technique. In addition, Sections 2.2 and 2.3 describe the applied Web IR techniques, namely the per-field normalisation and the evidence from the document URL path, respectively.

2.1 Language Specific Stemming

Stemming has been shown to be effective for ad-hoc retrieval from monolingual collections of documents in European languages [7]. Our main research hypothesis in this work is that applying the correct stemmer to a document and a topic would increase the retrieval effectiveness of the search engine. This would emulate a monolingual IR system, where the correct stemmer for both the language of the documents and topics is always applied.

To test our hypothesis, we use three approaches for processing the text of documents and topics. First, we do not use stemming. Second, we apply Porter's English stemmer to all text, regardless of the language. This approach stems English text, but it does not affect texts written in other languages, with the exception of those terms that contain affixes expected in English terms. Third, we identify the language of each document or topic, and then apply an appropriate stemming approach. The latter technique is called *language specific stemming*.

For identifying the language of documents or topics, we primarily use the TextCat language identification tool [2,15]. However, as the language identification process is not precise – often giving multiple language choices – we choose to supplement document language detection with additional heuristics. For each document, we examine the suggested languages provided by the language identification tool, and look for evidence to support any of these languages in the URL of the document, the metadata of the document, and in a list of "reasonable languages" for each domain. For example, we do not expect Scot Gaelic or Welsh documents in the Web sites of the Hungarian government.

For each of the considered languages, in addition to a stemmer, we assume that there exists an associated stopword list. The language of the queries is identified by either the provided metadata, or the TextCat tool.

2.2 Per-Field Normalisation

The evaluation of Web IR systems with known-item finding queries suggests that using the title of Web documents and, in particular, the anchor text of the incoming hyperlinks results in improved retrieval effectiveness [3,4].

In this work, we take into account the fields of Web documents, that is the terms that appear within particular HTML tags, and introduce one more normalisation method in the Divergence From Randomness (DFR) framework, besides *Normalisation 2* and the normalisation with Dirichlet priors [6]. We call this method *per-field normalisation*. Our introduced method applies term frequency normalisation and weighting for a number of different fields. The per-field normalisation has been similarly applied in [16] using the BM25 formula. In this work, we use a different document length normalisation formula.

Per-field normalisation is useful in a Web context because the content, title, and anchor texts of Web documents often have very different term distributions. For example, a term may occur many times in a document, because of the document's verbosity. On the other hand, a term appearing many times in the anchor text of a document's incoming hyperlinks represents votes for this document [5]. Moreover, the frequencies of terms in the document titles are distributed almost uniformly. Thus, performing normalisation and weighting independently for the various fields allows to take into account the different characteristics of the fields, and to achieve their most effective combination.

The DFR weighting model PL2 is given by the following equation:

$$score(d,Q) = \sum_{t \in Q} \frac{qtfn}{tfn+1}\Big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn)\cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn)\Big) \quad (1)$$

where $score(d,Q)$ corresponds to the relevance score of document d for query Q, $\lambda = \frac{F}{N}$ is the mean and variance of a Poisson distribution, F is the total term frequency in the collection, and N is the number of documents in the collection. $qtfn$ is the normalised query term frequency, given by $qtfn = \frac{qtf}{qtf_{max}}$, where qtf is the query term frequency, and qtf_{max} is the maximum query term frequency among the query terms. tfn is given by *Normalisation 2*:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg\_l}{l}), \quad (c > 0) \quad (2)$$

where tf is the frequency of t in a document d, c is a hyper-parameter, avg_l is the average document length in the collection, and l is the length of d.

Our per-field *Normalisation 2F* extends *Normalisation 2*, so that tfn corresponds to the weighted sum of the normalised term frequencies for each used field f:

$$tfn = \sum_f \Big(w_f \cdot tf_f \cdot \log_2(1 + c_f \cdot \frac{avg\_l_f}{l_f})\Big), \quad (c_f > 0) \quad (3)$$

where w_f is the weight of field f, tf_f is the frequency of t in f, avg_l_f is the average length of f in the collection, l_f is the length of f in a particular document, and c_f is a hyper-parameter for each field f. Note that *Normalisation 2* is a special case of *Normalisation 2F*, when the entire document is considered as one field. After defining *Normalisation 2F*, the DFR weighting model PL2 can be extended to PL2F by replacing tfn from Equation (3) in Equation (1).

2.3 Evidence from URL Path Length

In our previous work [11], we found that taking the length of the document URL path component into account is particularly effective in both topic distillation, and home page finding tasks. In particular, the relevance score of the retrieved documents is updated according to the following formula:

$$score(d,Q) := score(d,Q) \cdot \frac{1}{\log_2(1 + URLPathLen_d)} \quad (4)$$

where $score(d, Q)$ is the relevance score of document d for query Q. The length in characters of the document URL path component is denoted by $URLPathLen_d$. The above formula is applied for a certain number of top ranked documents in order not to introduce noise and not to change the ranking considerably.

In this work, we follow [16] and refine the combination of content analysis and evidence from the document URL path, by adding a score related to the length of the document URL path:

$$score(d, Q) := score(d, Q) + \omega \cdot \frac{\kappa}{\kappa + URLPathLen_d} \tag{5}$$

where ω and κ are free parameters. The parameter κ controls the saturation related to the length of the document URL path. The parameter ω weights the contribution of the document URL path length to the document's relevance score. We apply the above formula to all the retrieved documents.

3 Experimental Setting and Runs

In this section, we present how the above described techniques have been applied for the monolingual task of the WebCLEF 2005. More specifically, we provide details about our experimental setting (Section 3.1) and our submitted runs (Section 3.2).

3.1 Experimental Setting

For all our experiments we used a version of the University of Glasgow's Terrier platform. More details about the platform can be found in [10]. Terrier provides a range of weighting models, from classical models, such as tf-idf and BM25, to models based on the Divergence From Randomness (DFR) framework [1].

To support retrieval from multilingual document collections, such as the EuroGOV [13], it is essential that the IR system accurately and uniquely represents each term in the corpus. To meet this requirement, the correct encoding for each document must be identified prior to indexing. The version of the Terrier platform we used detects the encoding of documents. During the parsing of the collection, we used heuristics, based on the HTTP headers, the META tags, and the TextCat language identifier tool, to determine the correct content encoding for each document. Once the correct encoding is determined, each term is read and converted to UTF-8 encoding. This ensures that we have a correct representation of the documents.

For the language specific stemming technique, we mainly used the Snowball stemmers [14], with the following exceptions: English, where we used Porter's English stemmer; Icelandic, where we used the Danish Snowball stemmer; Hungarian, where we used Hunstem [8]; and Greek, where we did not apply any stemming. The anchor text of the incoming hyperlinks of documents is processed with the stemmer for the language of the source document. For example, when English documents link to a French document, then the anchor text of the French document is stemmed with Porter's English stemmer.

For the per-field normalisation, in all our experiments, we considered three document fields: the content, the title, and the anchor text of its incoming hyperlinks. The values of the hyper-parameters c_f and the weights w_f were automatically set using an extension of our previous work [6], which takes fields into account. The parameters ω and κ employed in the integration of evidence from the document URL path were empirically set using the .GOV TREC Web test collection and the associated known-item finding task from TREC 2003 [3].

3.2 Runs

As described in Section 2.1, we applied several stemming approaches to index the EuroGOV collection. Three indices were built: in the first one, no stemming was applied; in the second one, Porter's English stemmer was applied for all documents; and in the third index, the stemmer deemed appropriate for each document was applied.

We submitted five runs to the monolingual task of WebCLEF 2005, four of which used topic metadata. For all metadata runs, we used the domain topic metadata to limit the URL domain of the returned results. For example, if the topic stated `<domain domain="eu.int"/>`, only results with URLs in the eu.int domain were returned. For two runs, we also used the topic language metadata for detecting the correct stemmer to apply for the query. The official runs we submitted are detailed below:

- `uogSelStem`: This run did not use any metadata. Instead, we used the TextCat language identifier tool [2,15], to identify the language of each topic. The topic was then stemmed using the appropriate stemmer for that language. If TextCat was unable to classify the topic, then the topic was stemmed with the English Porter stemmer. We used the index with language specific stemming. This run tested the accuracy of the language identifier in determining which stemmer to apply to each topic.
- `uogNoStemNLP`: This run used only the domain metadata described above. No stemming was applied neither to the topics, nor to the documents. Additionally, we used a natural language processing technique to deal with acronyms. This run tested the retrieval effectiveness of not applying stemming in this multilingual Web IR setting.
- `uogPorStem`: This run used only the domain metadata described above. Porter's English stemmer was applied to all topics, and the corresponding index was used. This run tested the retrieval effectiveness of applying Porter's English stemmer to all languages in the EuroGOV collection.
- `uogAllStem`: This run used both the domain metadata described above, and the topic language metadata, which allowed the use of the correct stemmer for each topic. We used the index with language specific stemming. This run tested the hypothesis that applying the correct stemmer to both documents and topics would improve results overall.
- `uogAllStemNP`: This run is identical to `uogAllStem`, except that term order in the topics was presumed to be important. We applied a strategy where

the weights of query terms linearly decrease with respect to their position in the query and the query length. The underlying hypothesis is that in Web search, the user will typically enter the most important keywords first, then add more terms to narrow the focus of the search.

The used hyper-parameter and weight values related to the per-field normalisation are shown in Table 1. The used values for the parameters ω and κ are 2.0 and 18.0, respectively, for all the submitted runs.

Table 1. The used values of the hyper-parameters c_c, c_a, c_t, and the weights w_c, w_a and w_t, related to the per-field normalisation of the content, anchor text, and title fields, respectively, for the submitted runs

Run	c_c	c_a	c_t	w_c	w_a	w_t
`uogSelStem`	3.00	100	100	1	40	35
`uogNoStemNLP`	4.10	100	100	1	40	40
`uogPorStem`	3.19	100	100	1	40	40
`uogAllStem`	3.00	100	100	1	40	35
`uogAllStemNP`	3.00	100	100	1	40	35

4 Results and Discussion

Table 2 details the mean reciprocal rank (MRR) achieved by each of our submitted runs in the monolingual task. From initial inspection of the evaluation results, the run `uogNoStemNLP`, which did not apply any stemmers, gives the best MRR, closely followed by the run `uogPorStem`.

Table 2. Mean Reciprocal Rank (MRR) of the submitted runs to the monolingual task. The bold entry indicates the most effective submitted run, and the emphasised entry corresponds to the run without metadata.

Run	Description	MRR
`uogSelStem`	PL2F, URL, Language Specific Stemming	*0.4683*
`uogNoStemNLP`	PL2F, URL, No Stemming, Metadata, Acronyms	**0.5135**
`uogPorStem`	PL2F, URL, Porter's English Stemmer, Metadata	0.5107
`uogAllStem`	PL2F, URL, Language Specific Stemming, Metadata	0.4827
`uogAllStemNP`	PL2F, URL, Language Specific Stemming, Metadata, Order	0.4828

In Table 3, we have also broken the MRR down into the component languages of the queries, and the home page (HP) and named page (NP) queries. It would appear that Porter's English stemmer is more effective than, either no stemming, or the appropriate Snowball stemmer for Dutch and Russian. English and Portuguese topics give the best performance without any stemming applied. The weighting of the query term ordering showed little retrieval performance

Table 3. Mean Reciprocal Rank (MRR) of the submitted runs to the monolingual task. The bold entries indicate the most effective run for the corresponding set of topics. The numbers in brackets correspond to the number of queries for each language and each type of query. NP and HP stand for the named page and home page finding topics, respectively.

Topic Set		uogSelStem	uogNoStemNLP	uogPorStem	uogAllStem	uogAllStemNP
All	(547)	0.4683	**0.5135**	0.5107	0.4827	0.4828
DA	(30)	0.5168	0.5246	0.5098	**0.5857**	0.5829
DE	(57)	0.4469	0.4414	0.4567	**0.4780**	0.4689
EL	(16)	0.2047	0.3704	0.3659	0.3586	**0.4003**
EN	(121)	0.4988	**0.5578**	0.5240	0.5188	0.5239
ES	(134)	0.4198	0.4571	0.4635	0.4602	**0.4647**
FR	(1)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
HU	(35)	0.2713	**0.5422**	**0.5422**	0.1142	0.1003
IS	(5)	**0.3400**	0.3222	0.3222	**0.3400**	**0.3400**
NL	(59)	0.6362	0.6226	**0.6551**	0.6444	0.6447
PT	(59)	0.5262	**0.5565**	0.5336	0.5048	0.5028
RU	(30)	0.4838	0.4724	**0.4975**	0.4838	0.4625
NP only	(305)	0.4803	**0.5353**	0.5232	0.4952	0.4956
HP only	(242)	0.4531	0.4862	**0.4949**	0.4669	0.4666
All - HU	(512)	0.4818	**0.5116**	0.5085	0.5078	0.5089

improvement. It was particularly effective for the Greek topics (0.3586 to 0.4003), but showed very little positive or negative change for most languages.

The runs with the correct stemming applied (`uogAllStem` and `uogAllStemNP`) perform very well, with the exception of the Hungarian queries, which are affected considerably. The last row of Table 3 displays the MRR of all runs with all Hungarian topics removed. This shows that stemming makes little difference – the runs `uogAllStem` and `uogAllStemNP` achieve approximately the same MRR as the run `uogPorStem`, and are comparable to the run `uogNoStemNLP`.

The obtained performance when applying the correct stemmer to the Hungarian topics (the runs `uogAllStem` and `uogAllStemNP`) implies that the use of an aggressive stemmer, such as Hunstem [8], which addresses both inflectional and derivational variants, is not appropriate for the tested settings. However, when the language identifier classifies the Hungarian topics (as in `uogSelStem`), performance improves (0.2713 vs. 0.1142).

By comparing the runs `uogAllStem` and `uogSelStem`, we can see that the accuracy of the language classifier has an impact on the retrieval effectiveness (0.4827 vs. 0.4683 from Table 2). However, the effect of the classification accuracy for individual languages varied. For Hungarian topics, when the language identifier did not correctly classify the language of the topics, performance actually improved. The TextCat tool correctly classified 304 out of the 547 topics, while there were only 6 topics for which the identified language was wrong. TextCat did not classify 237 topics, because the input data was not sufficient to make a classification. For these topics, the Porter stemmer was used. Improvement in

MRR is obtained if no stemming is applied to the unclassified topics and the unstemmed index is used (0.4735 vs. 0.4683 from run uogSelStem).

We examined the benefit from using the field-based weighting model PL2F, as well as evidence from the URLs of Web documents (Sections 2.2 and 2.3, respectively). First the documents are represented by the concatenation of the content, title and anchor text fields, and the weighting model PL2 (Equation (1)) is used for retrieval. The hyper-parameter c of the *Normalisation 2* (Equation (3)) is set equal to c_c from run uogNoStemNLP (Table 1). The evaluation shows that this approach seems to be less effective than field-based retrieval with PL2F (0.5018 vs. 0.5135 from the run uogNoStemNLP). If the evidence from the URLs of Web documents is not used for the run uogNoStemNLP, then the obtained MRR is 0.5116, which is slightly lower than 0.5135 obtained from the run uogNoStemNLP. Overall, the field-based weighting model PL2F seems to have a positive impact on the retrieval effectiveness. However, the evidence from the URLs resulted in only a small improvement in retrieval performance.

We also investigated the average topic length, in particular for the German, Spanish, and English topic sets, and found these to be 3.3, 6.3, and 5.7 terms, respectively. In contrast, a recent study by Jansen & Spink [9] found an average length of 1.9, 2.6, and 5.0 terms for German, Spanish, and English queries, respectively. This difference can be due to two reasons. First, the studied queries in [9] are likely to include informational queries [12], which tend to be shorter, thus resulting in a lower average query length. Second, the difference in the average query length could be attributed to the fact that the used topics in WebCLEF 2005 were not representative of real European user search requests on a multilingual collection. Moreover, it's worth noting the distribution of queries in Table 3, where the number of queries by language in fact reflects the participating groups in WebCLEF 2005. Indeed, the creation of the queries and the corresponding relevance assessments was a joint community effort. In the future, it would be interesting to employ topics corresponding to European user search requests from commercial search engine query logs.

Regarding the two types of queries, all our submitted runs performed consistently better on the named page finding queries, than on the home page finding queries. Overall, all our submitted runs to the monolingual task of WebCLEF 2005 were clearly above the median of all the participants' submitted runs. Four of our runs were the best performing runs overall, and our run `uogSelStem` was the best performing run among the compulsory runs without metadata.

5 Conclusions

Our participation in the WebCLEF 2005 has been focused on the correct application of stemmers in a multilingual setting, as well as on the use of different document fields and evidence from the document URL path. We found that applying the correct stemmer for the language of the document and topic was effective in most cases. However, the improvements in retrieval performance from applying the correct stemmer for a language depend on the accuracy of the

language identification of topics and documents. Without accurate language identification, retrieval effectiveness is penalised when a different stemmer is applied to a topic and the corresponding target document. The bare-system approach of applying no stemming at all achieved the best performance in the monolingual task.

Regarding the Web IR techniques we used, the per-field normalisation seemed to improve the retrieval performance, while the document URL path length resulted in smaller improvements. In future work, we will be extending per-field normalisation to other common fields of Web documents, such as H1, H2 tags.

References

1. G. Amati. *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, Dept of Computing Science, University of Glasgow, 2003.
2. W.B. Cavnar and J.M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94*, pp. 161–175, 1994.
3. N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the TREC-2003 Web Track. In *Proceedings of TREC 2003*, 2003.
4. N. Craswell, D. Hawking. Overview of the TREC-2004 Web Track. In *Proceedings of TREC 2004*, 2004.
5. D. Hawking, T. Upstill and N. Craswell. Towards better weighting of anchors. In Proceedings of ACM SIGIR 2004, pp. 512–513, 2004.
6. B. He and I. Ounis. A study of the Dirichlet Priors for term frequency normalisation. In Proceedings of ACM SIGIR 2005, pp. 465–471, 2005.
7. V. Hollink, J. Kamps, C. Monz and M. de Rijke. Monolingual Document Retrieval for European Languages. *Information Retrieval*, 7(1-2):33–52, 2004.
8. Hunspell & Hunstem: Hungarian version of Ispell & Hungarian stemmer. URL: `http://magyarispell.sourceforge.net/`.
9. B.J. Jansen and A. Spink. An analysis of Web searching by European AlltheWeb.com users. *Inf. Process. Manage.*, 41(2):361–381, 2005.
10. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. Terrier Information Retrieval Platform. In *Proceedings of ECIR 2005*, LNCS vol. 3408, pp. 517–519, 2005. URL: `http://ir.dcs.gla.ac.uk/terrier/`.
11. V. Plachouras, B. He, and I. Ounis. University of Glasgow at TREC2004: Experiments in Web, Robust and Terabyte tracks with Terrier. In *Proceedings of TREC 2004*, 2004.
12. D.E. Rose and D. Levinson. Understanding user goals in web search. In Proceedings of the 13th international conference on World Wide Web, pages 13–19, 2004.
13. B. Sigurbjonsson, J. Kamps, and M. de Rijke. EuroGOV: Engineering a Multilingual Web Corpus. In *Proceedings of CLEF 2005 Workshop*, 2005
14. Snowball stemmers. URL: `http://snowball.tartarus.org/`.
15. G. van Noord. TextCat language guesser. URL: http://odur.let.rug.nl/~vannoord/TextCat/.
16. H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft Cambridge at TREC-13: Web and HARD tracks. In *Proceedings of TREC 2004*, 2004.